DOCUMENT RESUME

ED 279 688 TM 870 075

AUTHOR Miller, George A.
TITLE Explanatory Skills.
PUB DATE 86
NOTE 33p.; One of 46 papers commissioned by the Study Group on the National Assessment of Student Achievement and cited in Appendix B to their final report "The Nation's Report Card" (TM 870 049). For other papers in this group, see TM 870 050-094.
PUB TYPE Viewpoints (120) -- Information Analyses (070)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Cognitive Processes; Difficulty Level; *Educational Assessment; Educational Testing; Elementary Secondary Education; *Inquiry; Learning Theories; Mathematical Logic; *Mathematics Tests; National Surveys; *Problem Solving; *Science Tests; *Scientific Literacy; Test Format; Testing Problems; Testing Programs
IDENTIFIERS *National Assessment of Educational Progress; Thinking Skills

ABSTRACT

In assessing the quality of science teaching for an effort such as the National Assessment of Educational Progress (NAEP), it is important to understand what is meant by scientific thinking--the search for explanations. Instruction should involve higher-order cognitive skill development, but it is difficult to measure reasoning and understanding with traditional multiple choice tests. Additionally, time spent on improving students' multiple-choice test performance reduces the time available for strengthening more complex skills. Increased use of tests in other formats would encourage the teaching of these skills. Cognitive theorists have recognized that thinking and learning rely heavily on the construction of explanations. In the many examples of learning activities presented in this paper, cognitive learning is shown to be driven by the search for explanations. Current developments in cognitive research justify that effort be dedicated to the assessment of explanatory skills. Testing alternatives include multiple choice tests designed with less emphasis on factual memory, essays, free response tests, interviews, thinking aloud protocols, and performance tests. Empirical studies of explanation are recent. They include literature on the development of explanatory skills and on the social role of explanations. Tests of divergent productions are the result of psychometric studies of explanation. The development and use of tests that sample students' ability to offer scientific explanations could add an important dimension to NAEP. (GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED279688

Explanatory Skills

George A. Miller

Princeton University

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☑ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Paper commissioned by

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

G. A. Miller

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

THE STUDY GROUP ON THE NATIONAL ASSESSMENT OF STUDENT ACHIEVEMENT

1986

TM 810 075

BEST COPY AVAILABLE

Explanatory Skills

George A. Miller
Department of Psychology
Princeton University

A highly technological society depends on scientists and engineers; it must be supported by an educational system that can provide a steady supply of men and women able to think scientifically. It is a matter of some importance, therefore, to determine whether this expectation is being met—whether in fact American schools are training students to think like scientists and mathematicians.

Which raises a difficult question: How can the quality of science teaching be evaluated? The problem, of course, is that in order to make such assessments it must be possible to say with some precision what it means to think like a scientist. That is not a simple task, and this working paper can address only one component of it. This paper attempts to argue that an important part of any scientist's thinking is a *search for explanations*.

In 1961, in his introduction to *The Structure of Science*, Ernest Nagel wrote: "It is the desire for explanations which are at once systematic and controllable by factual evidence that generates science; and it is the organization and classification of knowledge on the basis of explanatory principles that is the specific goal of the sciences" (1961, p. 4). Nagel was not making some extravagant claim. He was simply reminding his readers of something that he expected any sensible person would take for granted, thus establishing a common ground from which his more arguable explorations into the nature of science could begin.

It is noteworthy that this commonplace observation about the origin and goal of modern science has received little attention from those whose responsibility it has been to evaluate science teaching. It would be unreasonable to criticize those psychometricians too severely, however, because they have not been alone in their neglect of this important cognitive process. Indeed, it is only recently that the importance of

explanatory processes has been explicitly recognized by cognitive scientists, many of whom are now beginning to suspect that the desire for explanations generates far more than science (Schank, in press). Work aimed at revealing the source and patterning of explanations is now being actively pursued in several university laboratories.

This rediscovery of the obvious has important implications for instruction, some of which are presently being explored; this working paper will attempt to push the implications beyond research and instruction into the arena of assessment. But since this proposal—the proposal to evaluate the quality of science and mathematics teaching by sampling students' ability to give scientific and mathematical explanations—is still in its infancy, this paper will begin at the beginning: with a discussion of why this reasonable proposal has not yet been adopted, and with some arguments why it should be adopted now.

Some Basic Assumptions

> *Teaching is not, and should not be, simply a matter of conveying information for students to remember. Good teachers help their students in many ways, one of the most important being to show them how to use their minds more effectively—how to think clearly, critically, and coherently.*

This premise should be obvious—it should not be necessary to state it explicitly. It should be possible to presuppose the educational importance of higher cognitive skills, and to move directly to such pressing questions as how these skills should be characterized, how they can be taught, and how the success of this teaching can be evaluated fairly and objectively. Unfortunately, however, the remarkable success of standardized tests in education created a situation in which it became necessary to reaffirm fundamental educational goals.

Briefly, this was the situation: First, the development of objectively scored, multiple-choice achievement tests made standardized, mass testing possible. The results of such tests were then used, directly or indirectly, to assess the quality of education; not only were test results widely publicized, but important policy and budgetary decisions were based on them. Consequently, teachers were pressed to "teach to the

tests," to teach their students how to score well on what was generally accepted as the true measure of educational achievement.

The desire to hold schools publicly accountable is certainly proper; this method of doing so would also have been proper if standardized, multiple-choice tests actually measured what everyone agrees that children ought to learn. Experience with such testing, however, provided convincing evidence that good performance with a multiple-choice format depends primarily on good recognition memory, and does not adequately sample other cognitive skills of equal, if not greater, importance. When teachers concentrate on improving their students' performance on multiple-choice tests, therefore, it reduces the time spent strengthening more complex, or higher-order skills.[1]

To neglect higher-order cognitive skills can have unfortunate effects. By their nature, these skills organize thinking into effective channels; without them, no degree of achievement in simple cognitive skills can ensure coherent, goal-directed action. Speed in adding numbers is of little value unless you know what numbers should be added; accuracy in spelling if useless if you have nothing worth writing; memory for facts will not help you decide which facts are relevant to your problem. Simple skills are necessary but not sufficient. To the extent that the effect of mass testing was to emphasize the accumulation of simple skills at the expense of more complex processes of reasoning, judgment, planning, and management, a serious distortion had occurred.

As early as 1962 the mathematician Banesh Hoffmann deplored the widespread emphasis on multiple-choice tests in educational research and attempted to rally public opinion against what he considered to be their superficiality. Lacking any better methods of assessment, however, leaders in the field of educational testing were at first reluctant to admit the dangers inherent in this situation. Fortunately, their resistance has now lessened. For example, in 1984 Norman Frederiksen, one of the intellectual

[1] Different authors give different definitions of "higher-order skills" depending on context. Even lacking a precise definition, however, most people can recognize them when they see them. Resnick (1986) lists some attributes that make recognition possible: they are non-algorithmic, are complex, have multiple solutions, require nuanced judgment, satisfy multiple criteria, involve uncertainty, are self-regulated, impose meaning on disorder, are effortful.

leaders of the Educational Testing Service, wrote:

> "There is evidence that tests do influence teacher and student performance and that multiple-choice tests tend not to measure the more complex cognitive abilities. The more economical multiple-choice tests have nearly driven out other testing procedures that might be used in school evaluation. It is suggested that the greater cost of tests in other formats might be justified by their value for instruction—to encourage the teaching of higher-level cognitive skills and to provide practice with feedback" (p. 193).

This summary statement of what Frederiksen calls "the real test bias" provides a starting point for this working paper.

Efforts to change this situation must be continued and strengthened, and an important step in that direction is to develop and use testing instruments that can assess complex skills.

> *The measurement of higher-order cognitive skills is essential if we are to determine the extent to which the goals of education are actually being met.*

Measuring instruments are needed to test the whole range of skills that the educational system is expected to impart. NAEP should be particularly alert to the limitations of traditional multiple-choice tests and to the need for new measurement techniques.

Simply abandoning a multiple-choice format will not solve the problem, however. Any assessment technique that emphasizes rote memorization runs the risk of weighting simple cognitive skills too heavily at the expense of more complex processes of reasoning and understanding. The mathematician Alan Schoenfeld (1985) has reported his observations of a Tenth Grade geometry class preparing to take the New York State Regents examination. The teachers knew that there are about a dozen required constructions and that one or two of them (each worth 2 points out of 100) would appear on the examination. As a result, students were expected to memorize those constructions and practice them until they could perform them rapidly without error. The rationale for a given procedure was presented before the procedure was practiced, but the importance of that rationale was forgotten during the practice sessions.

Schoenfeld comments that "A very strong classroom emphasis on performance—on memorizing constructions and practicing them until they can be performed mechanically with a very high degree of accuracy—ultimately results in students losing sight of the rational reasons for the correctness of those constructions" (p. 374). Moreover, it left most students with a false impression of what mathematics is all about.

Simple fiddling with test formats, therefore, is not likely to solve this problem. What is needed are methods of testing that will probe beyond memory into those higher cognitive processes that the schools are expected to strengthen. In order to develop such methods, we need a clear conception of what those higher-order cognitive skills are and how they can be measured. Fortunately, a theory of cognitive skills is rapidly taking shape at the present time.

> *The measurement of higher-order cognitive skills must be based on the best cognitive theories of learning and reasoning that are available.*

In order to indicate what is meant by the best cognitive theories available, some examples will be discussed—examples chosen with with an eye to their special relevance for explanation-based cognitive processes.

Advances in Cognitive Theory

The past two decades have seen a revolution in our understanding of cognitive processes and our ability to simulate them with intelligent automata. This progress has been paralleled by an active public interest in the teaching of thinking and a willingness on the part of educators to accept innovations suggested by the new cognitive theories. Robert Glaser (1985) summarized this situation:

> "This confluence of renewed educational interest and modern scientific investigation offers a challenging opportunity to attack the problem of 'passive knowledge'—knowledge that students receive and express, but cannot use effectively for thinking and learning. Schools have not been well enough equipped for this task. . . . Today, however, educators, educational researchers, developmental psychologists, and cognitive scientists are designing school programs and conducting investigations on understanding and problem solving in

> mathematics and science, comprehending and reasoning with text material, study skills and ability to learn, and the role of memory organization in the acquisition of knowledge. These processes of human cognition and learning are being studied with particular attention to how conditions that foster them might be built into the materials, methodologies, and environments for learning and schooling" (p. ix-x)

Accounts of these developments can be found in Segal, Chipman, & Glaser (1985), Nickerson, Perkins, & Smith (1985), and Resnick (1986). The time is ripe to develop better methods to assess the educational value of these innovations. NAEP can play an important role in fostering this development.

Progress in understanding higher cognitive processes has benefitted enormously in recent years from the growth of cognitive psychology and its technological ally, artificial intelligence. Indeed, it is from this work on the psychological nature of knowledge representations and transformations that the view of thinking as a skill, or collection of skills, has received the strongest support. Any serious attempt to describe what a person engaged in productive thinking is doing must recognize immediately that a variety of distinguishable activities are being coordinated hierarchically to achieve the intended goal.

When thinking was first simulated in computer programs the methods that were used to solve problems were very general—planning, means-ends analysis, backward chaining, and the like—with little or no reliance on detailed knowledge of the problem domain. These heuristic methods, which were pioneered by the mathematician George Polya (1957), can be applied to a wide range of problems; versions of them "can be taught in such a way that they can be used in new domains where they are relevant" (Simon, 1980, p. 86; but see Schoenfeld, 1985, ch. 3).

In complex problem domains, however, these so-called "weak heuristics" are often insufficient. Detailed knowledge of the domain is often required. The "expert systems" that were subsequently developed placed much greater emphasis on the knowledge base, sometimes at the expense of more general problem-solving methods. The attempt to strike an appropriate balance between general heuristics and domain-

specific expert knowledge offers an interesting parallel to the assessment problem: how to strike an appropriate balance between higher thinking skills and simpler factual and procedural information.

Many aspects of cognitive theory were refined or reformulated in the course of these developments. Of particular relevance for the present discussion were certain proposals concerning the cognitive bases for logical thinking. Initially it was assumed that deductive logic would provide an appropriate theoretical framework within which to study human reasoning processes, and that inductive logic would provide an appropriate framework for theories of learning. Neither of these assumptions has proved productive.

Deductive logic is used to argue from initial premises through intermediate steps to some final conclusion that is implied by the premises. By *modus ponens*, for example, one is allowed to argue that, given *P* and *If P, then Q*, it follows that *Q*. But when attempts are made to formulate actual situations in these terms, they may fail. For example (Harman, 1986), if Mary believes *If I look in the cupboard, I will see a box of Cheerios* and *I am looking in the cupboard*, then by *modus ponens* she should also believe *I see a box of Cheerios*. But if, in fact, Elizabeth finished the Cheerios yesterday and the box is now gone, Mary should not conclude that she is seeing a box of Cheerios that isn't there! Instead, she should revise her premises. Practical reasoning follows different principles from the rules of deductive argument.

Similarly, the rules of inductive logic proved to be less helpful in theories of learning than was originally hoped. The impossibility of proving a general statement on the basis of a finite number of positive instances had long been recognized, but the whole theory of induction was dealt a heavy blow when Nelson Goodman (1955) pointed out that an indefinite variety of (incompatible) hypotheses could all be confirmed equally by the same observation. That is to say, the "inductive leap" from a finite set of examples to a universal generalization can land almost anywhere. It should not have been surprising, therefore, when cognitive scientists realized that different people can draw very different conclusions after observing the same set of instances—that the consequences of experience are as much constructed by as imposed

on the learner.

As these difficulties became increasingly obvious, alternatives to the classical theories were put forward.

Explanation-Based Theories

Many cognitive theorists now take as fundamental the assumption that thinking and learning are manifestations of a general human search for meaning. This search relies heavily on the construction of explanations. For example, in their studies of people learning to use word processors, Clayton Lewis and his colleagues (Lewis & Mack, 1982; Mack, Lewis, & Carroll, 1983) noticed that learners frequently offered explanations (usually spurious) of why things had happened the way they did. Their learners were not told to give explanations, yet they showed considerable fluency and ingenuity in formulating them. Lewis (1986) speculated that these spontaneous explanations can help in the formation of useful generalizations: determining how their actions are related to observed outcomes enables learners to build new procedures for performing novel tasks. This line of argument leads to a theory of generalization very different from traditional theories of inductive learning.

One criticism of explanation-based theories is that people—especially novices—so often come up with absurd explanations. What use are absurd explanations? The answer seems to be that explanations either do matter or they don't. If the explanation makes a difference for what the person is doing, its absurdity will quickly become obvious and it can be discarded. If the explanation does not matter—if the explanandum never recurs—then little harm will be done by allowing it to go uncorrected. Explanations that cannot connect up with a person's other beliefs are probably soon forgotten.

It should be noted that, as the term "explanation" is normally used, there are two aspects to an explanatory episode. One aspect is the formulation of hypotheses about the reasons for some state or event, and the selection of one hypothesis over the others; the second is the expression of the selected hypothesis in a form that others can understand or be persuaded to accept. The expressive side of explanation has received

some attention from developers of expert systems: it seems that human users of expert systems are often reluctant to accept a computer's decision without some explanation of the computer's reasons for that decision. In the present discussion, however, attention is focussed on the first aspect, the formulation and selection of explanatory hypotheses or, loosely, how people explain things to themselves.

It should be kept in mind that the correct formulations of cognitive theories of concept learning and practical reasoning are still under debate; the views summarized here are interesting more as examples of the current state of the field than as final answers to these old and difficult questions.

Explanation-Based Comprehension. The first step toward explanation-based theories of cognition was the recognition that comprehending linguistic messages is a special kind of problem solving. At first the problem was assumed to be one of discovering logical links between successive sentences, but then a kind of complementarity between comprehending messages and solving problems was observed.

A message source produces sounds or makes marks; a receiver observes these signals and tries to infer what mental state would lead the source to produce them (Miller, 1979). In simpler terms, comprehending a message is a matter of explaining the source's communicative behavior. Moreover, if the message describes the activities of other people, that part of the message will be understood by constructing explanations for why they acted as they did.

This insight might have come earlier but for the fact that comprehending is problem solving in reverse. As Wilensky (1983) puts it:

> "Problem solving has often been abstractly characterized as searching through a 'solution space' for an answer to some problem. That is, given a goal, a set if operators, and an initial state of the world, the task is to construct a sequence of operators that transforms the initial state into the goal state. Understanding involves just the opposite in terms of what is known and what must be computed. . . . The reader is given the 'solution,' in problem-solving terminology, and must reconstruct the goal and state of the world from it. Rather than searching through solution space for an answer, and understander searches through an 'explanation

space' to find a set of circumstances that would explain a character's behavior'' (p. 10).

Prior to this reformuation, the problem of comprehending text had been viewed as a problem of finding relations between sentences. An example from Schank & Abelson (1977):

Willa was hungry. She picked up the Michelin guide.

The problem was to construct a sequence of logical steps connecting the second sentence to the first. Since that connection obviously involves information that is not given in either sentence, it must be drawn from memory. The problem of representing that memory was solved by introducing plans and scripts—stereotyped knowledge structures based on experience, in this case, experience in satisfying hunger.

Scripts are useful in providing connections (by default, if necessary), but they tend to be rigid and not easily adapted to nonstereotypical situations. Explanation patterns are more flexible (Kass, 1986). Replacing scripts by explanations yielded a more general theory of language comprehension. According to the explanation-based theory of comprehension, much of the work that a receiver needs to do in order to comprehend revolves around the task of finding explanations for what is said or written (Mooney & DeJong, 1985).

Once the advantages of this reformulation were appreciated, explanation was found useful in accounting for other cognitive phenomena. For example, Roger Schank (1981, 1982) proposed that they held the key to an experience he called "reminding"—when anomalous event *A* causes you to recall another, similar event *B*. In order to understand *A*, you construct an explanation similar to the explanation you previously constructed for *B*. One implication of this hypothesis is that people must have very good memories for explanations (Leake & Owens, 1986).

Recently the idea has been applied still more broadly. For example, Kristian Hammond (1986) has analyzed the explanation of failures as a way to anticipate and avoid pleanning problems. Explaining is ubiquitous. In Schank's words: "People want to understand the world—personally, socially, and physically. They do this by constantly creating and modifying explanations and indexing memories by the

explanations they cause to be formed" (1986, p. 141).

Explanation-Based Generalization. The idea that cognitive learning might be driven by a search for explanations has taken different forms (DeJong, 1981; Kedar-Cabelli, 1985). Tom Mitchell and his colleagues (Mitchell, Keller, & Kedar-Cabelli, 1986) have developed this view as a theory of generalization.

Suppose that, instead of searching through large bodies of data for similarities among instances, the learner already has a definition of the concept being studied, a general understanding of the domain, and a capacity for constructing explanations as to why some new instance is an example of that concept. Given this initial knowledge base, a great deal can be learned from just a single training example.

In this context, an explanation of how an instance is an example of a concept is an argument that the example satisfies the concept definition. For example, suppose the concept to be learned is a relation R between objects such that $R(A,B)$ holds between objects A and B iff it is safe to stack object A on top of object B. Suppose further that the learner's initial definition of R is that it is safe to stack A on B if A is lighter than B, and that the learner has a general understanding of size, volume, density, weight, etc., in terms of which explanations could be constructed,

Now imagine that a particular instance, $R(a,b)$, is presented. The first step is explanatory. The learner attempts to determine which features of the example are relevant for characterizing R. This task is accomplished by constructing an explanation of how this example satisfies the concept definition. In this case, the full explanation will be:

object a has weight W_a;

object b has weight W_b;

W_a is less than W_b;

therefore, object a is lighter than object b;

therefore, according to the definition of R, it is safe to stack a on b.

Thus, it has been explained why this example is a positive instance of the concept being studied. Note that the explanation tells us that only the weights of the two

objects are relevant to *R*, not their color, ownership, structure, or any other property.

If it turns out that the explanation is wrong, if *a* is not lighter than *b*, the initial definition will have to be revised—which may require real thinking, not simple learning from examples.

If the explanation succeeds in isolating relevant dimensions of the training example, a second step may try to refine the definition of the concept. If, for example, the weight of *a* is half that of *b*, the definition may be revised to say that *R(A,B)* holds if object *A* is lighter than half object *B*. The revised concept definition will typically be a specialization of the concept. How explanations can be generalized is presently an active research issue (DeJong, 1986; Kass, 1986; Mooney & Bennett, 1986).

Although explanation-based accounts of learning make relatively strong assumptions about a learner's initial state of knowledge, the reward is that they can account for the fact that learners often generate a justified generalization from a single training example. In that respect they seem superior to the classical theories of learning by induction of similarities among multiple instances, and are considered promising for the development of computer programs that can learn from exposure to training examples of a concept.

Explanation-Based Reasoning. Whereas only some kinds of learning can be said to be "explanation-based," it is not implausible to argue that all practical reasoning relies on explanatory skills. This view conflicts with the common assumption that reasoning must be deductive.

The two views of reasoning cannot be equated. There are explanations that are not deductive, as when a person's behavior is explained in terms of motivation or intention, or when probability is used to explain the likelihood of an outcome. And there are deductions that are not explanatory: every proposition implies itself, for example, but not every proposition explains itself. A less trivial example of a nonexplanatory deduction is the following:

That pole is 8 feet, 8 inches tall.

Its shadow is 5 feet long.

Therefore, (by trigonometric inference) the sun is at an angle of 60° in the sky. This deduction might serve as part of an explanation, but in itself it does not explain anything—it certainly does not explain the sun's being at that angle (Bromberger, 1966).

Gilbert Harman (1986) has developed the thesis that practical reasoning is a process that leads to a revision of the reasoner's belief structure—a change in the system of internal representations that comprises a person's view of the world and the actors in it. Revisions can occur by adding new beliefs, either through observation or reflection, or by subtracting old ones, as when beliefs are abandoned.

Revisions occur only in the service of some interest or goal: for example, an interest in answering a question, understanding what someone is up to, discovering how to discharge some obligation, etc. Reasoning is a response to that motivation, an attempt to satisfy the interest or achieve the goal by making small additions, revisions, or deletions in order to improve the explanatory coherence of the whole set of beliefs.

When you are considering whether or not to accept some given conclusion *P*, you must consider whether there is some explanatory structure that you can accept that includes *P*. You must be able to determine what acceptable explanatory structures are available. Moreover, if more than one explanation is available, you must be able to select the best one from among the competing alternatives. In general, the best explanation will be the one that accounts for the most data with the least change in your existing system of beliefs. Harman comments that "Belief revision is like a game in which one tries to make minimal changes that improve one's position. One loses points for every change and gains points for every increase in coherence" (p. 68). The process is subject to a number of constraints that need not be reviewed here.

The unit of reasoning is an explanatory structure. Successive steps in an explanatory structure are linked by immediate intelligibility, not necessarily by logical implication; in familiar domains, multi-link explanatory structures can become units or rules that guide successive steps in the reasoning process. Since there is a limit to the number of steps of reasoning can be actively considered at any one time, there is an obvious advantage to large steps.

Much remains to be done before the implications of this approach will be well understood. However, the theory of explanation-based reasoning has already been developed to the point that computer simulations of such reasoning can be created and explored (Harman, Cullingford, Bienkowski, Salem, & Pratt, 1986).

From these brief examples it should be obvious that the cognitive theories currently under development are much more complex than earlier theories were, and much closer to one's intuitive feeling for the mental operations that go on during learning and reasoning. Some of the implications of these theories for instruction are already being actively explored (see, for example, Segal, Chipman, & Glaser, 1985; Chipman, Segal, & Glaser, 1985), but beyond the obvious fact that such considerations greatly complicate the evaluation of a person's performance, relatively little psychometric research has been done on the related problem of assessment.

> *Explanation is sufficiently central to current developments in cognitive theory, and sufficiently representative of the kind of higher-order cognitive processes that should be strengthened by education, to justify a special effort dedicated to the assessment of explanatory skills.*

The real test of the new theoretical ideas about explanation-based cognition will be how well they conform to observational data on real persons struggling to explain real phenomena. Better methods of assessment can be expected to grow out of such observations. Before considering some of the studies that have been conducted, however, we should first review the kinds of observational techniques that are available for gathering the relevant data.

Alternative Methods of Data Collection

Methods of investigation that have been used to study students' knowledge of science and mathematics have ranged from multiple-choice tests to detailed transcriptions of students' thoughts as they have tried to solve problems. In order to determine whether students have acquired real expertise in any subject, the more elaborate methods of data collection are essential, but they are also expensive, time-consuming, and often require special skill to interpret.

The following list considers some of the advantages and disadvantages of the more popular investigative techniques.

Multiple-Choice. The great advantage of multiple-choice tests is that they can be scored rapidly and objectively. Before giving up that advantage, it would be prudent to consider whether the multiple-choice format might be preserved if less emphasis were placed on factual memory.

Experts in test construction generally agree that it is possible to write multiple-choice items that force students to think. Bert Green (1981) offered the following pair to illustrate the difference between items that test recognition memory and items that demand reflection and judgment. First, an item that tests memory:

The mean of a set of values is

(a) the lowest value

(b) the average value

(c) the middle value

(d) the most frequent value.

Any student who recognizes the definition that was given in the textbook can answer this item correctly. By contrast, consider:

The correlation of SAT-verbal or SAT-math among all test-takers is about .5. For the group of applicants admitted to Harvard, the correlation is probably

(a) greater than .5

(b) about .5

(c) less than .5

(d) anything—there is no basis for a guess.

Green comments that in order to select the correct alternative on this item: "The students will have to reason that, since Harvard is a very selective school, the admitted group will have relatively homogeneous test scores, so the correlation will be less than in the national group. Probing items can be written about almost any subject, given enough time and thought" (p. 1003). But he admits that such items are hard to construct. Consequently, most multiple-choice items sample memory for relatively superficial information. Various attempts to estimate the fraction of actual items that

test memory have put it at more than half.

Essay Questions. Over the years, the multiple-choice test's strongest competitor for measuring academic achievement has been the essay test, which requires students to compose a paragraph or more of intelligible prose. Writing essays has the advantage that it taps more complex cognitive processes; moreover, essay questions are relatively easy to prepare. But the costs of scoring—not just the expense and the time required, but the subjectivity of the scoring and the low reliability of the results—more than outweigh those advantages in the eyes of most professional testers.

Free-Response. Free-response items seem to offer a compromise. By requiring students to generate the answers, the free-response question seems to share some of the desirable properties of the essay question, yet avoids its subjectivity. That is to say, the short answer, usually a word or phrase, can be recognized and scored objectively. Yet it is possible for a free-response item to tap more complex cognitive processes because students are asked to generate correct responses, not merely recognize them.

As free-response items are generally used, unfortunately, this possibility has not been realized. Several studies have indicated that when multiple-choice tests are reformulated as free-response tests, the scores on the two are highly correlated. Most researchers interpret those results to mean that the two kinds of tests measure the same thing.

That interpretation can be challenged, however. Just as it is possible to construct multiple-choice items that require thinking, so it is possible to construct free-response items that involve thinking. Free-response items *could* measure different skills, even though they have usually been matched to multiple-choice items in such a way as to equate the two forms. Frederiksen (1984) has pointed out that high correlations between the two forms are obtained only when the free-response items are adapted directly from multiple-choice items—they usually have exactly the same stems, the only difference being that one tests recall memory and the other tests recognition memory. However, when students are asked to formulate a hypothesis, rather than simply recognize the best hypothesis on a list of possibilities, the correlation between free-response and multiple-choice tests can be very low (Ward, Frederiksen, &

Carlson, 1980).

Standardized Interviews. Jean Piaget is usually credited with having demonstrated the value of interview techniques for studying cognitive development, although considerable intuition went into his selection and interpretation of children's answers. Subsequent investigators have attempted to standardize their interviews: the same question is asked of a large number of children, their responses are categorized, and different, standardized prompts are formulated in order to follow up each category.

Although interviews are time-consuming and expensive to administer, they have the advantage of providing diagnostic information about the responder's thought processes. If the only information provided by a test item is that the answer is right or wrong, it is usually impossible to determine what processes led to that result. Often there is more than one way to arrive at a correct answer; different ways of solving a problem can be diagnostic of the solver's level of sophistication, and such differences are usually indicative of higher-order skills of planning and resource management. A carefully standardized interview can elicit such information, but the method is generally reserved for research, rather than assessment.

Thinking-Aloud Protocols. The fullest reports of the problem-solving processes are obtained when thinkers are trained to say aloud what alternatives they are considering at each stage in the solution of a problem. The problems studied by this method require considerable thought—they are not tests of factual memory—and may take many minutes to solve. Subjects usually work with minimal guidance from or interaction with the investigator.

Each subject's running report is recorded and later transcribed for careful analysis. It is not unusual to spend several hours analyzing and re-analyzing a single hour of recorded protocol. A good subject—someone who can think out loud unselfconsciously—can provide an amazingly detailed description of the steps taken in trying to solve the problem. Indeed, such protocols are sufficiently informative that computer programs can often be written to simulate the person's thought processes (Newell & Simon, 1972).

Protocol analysis would be far too expensive and labor-intensive for serious consideration as a psychometric instrument in survey research. It is mentioned here simply as a reminder of the richness of information that is available under ideal conditions, information that could never be uncovered by multiple-choice tests.

Performance Tests. Motor skills have usually been measured rather differently from cognitive skills: students are asked to demonstrate their level of skill in an actual "hands-on" performance. Such tasks have high fidelity to practical situations, and no special materials need to be prepared, pre-tested, and standardized. In order to standardize the results of such tests it is necessary to control the physical conditions under which the test is given and the method of administration and scoring of the test (Wigdor & Green, 1986).

When this approach is adapted to cognitive skills it can offer the advantage of assessing a student's performance in a situation closely resembling that in which the skill will be used. To test spelling, typing, or adding numbers, for example, the obvious recourse is to ask the student to spell, type, or add. Since not all practical situations are so easily simulated, however, cost and efficiency weigh against this kind of test.

This brief review of alternative test formats illustrates the general point that nothing is more efficient and economical than a multiple-choice test. Or, to put this conclusion the other way round,

> *Any attempts to assess higher cognitive processes adequately and objectively will necessarily be more expensive than the multiple-choice tests now in general use. However, the need to assess higher cognitive skills is sufficient to justify the additional expense of developing and administering new methods of testing.*

Empirical Studies of Explanation

Since the theoretical importance of explanatory skills has been recognized only recently, relatively few empirical studies of explanation have been published. However, the small literature that does exist is suggestive.

Development of Explanatory Skills. A basis for the empirical investigation of explanatory skills was laid down in some of Piaget's earlier works, particularly *The Child's Conception of the World* (1929) and *The Child's Conception of Physical Causality* (1960). Using his clinical interview technique, Piaget recorded children's explanations of a wide range of phenomena: the nature of dreams, the origin of the wind, explanations of mechanical devices, etc.

For example, a model steam engine whose workings were almost entirely visible was shown to children and their explanations were recorded: children under 6 years of age told him that the fire directly produced the movements of the wheels; between 6 and 8 years, the water in the boiler and a pipe connected to the wheel were said to be necessary; not till 8 or 9 years did they say that the fire and water combined to make steam, and that it was the steam that drove the wheel.

Leaving aside Piaget's claims about stages of cognitive development, his data certainly show a developmental trend. The trend can be characterized as one of increasing complexity—a normal consequence of learning. Piaget attributed animistic and magical explanations to the younger children, but their responses may have been due more to ignorance than to faulty reasoning (Carey, 1985). The more remarkable fact, from the present vantage point, is that even the youngest children seemed to understand the need for an explanation, even before they were prepared to give an adequate one.

Subsequent workers in the United States have revived Piaget's interest in children's thinking about physical phenomena and have attempted to conduct the interviews in a more rigorous and standardized manner. Work by Krupa, Selman, & Jaquette (1985) can be taken as representative of these more recent studies of children's explanations for scientific phenomena.

Consider how they constructed standardized interviews. First, they set up two demonstrations, one involving gravity and the other electromagnetism, which they showed to children selected from the First, Third, Fifth, Seventh, Ninth, and Eleventh Grades. Both problems allowed direct manipulation of the materials and posed at least one experiment whose outcome conflicted with the child's expectations. They

comment that "The resulting cognitive conflict increased the ease with which we could observe the child's knowledge as he tried to explain the phenomena" (p. 433).

They asked each child the same sequence of questions. For example, in the case of the electromagnet, its property is demonstrated by its effect on a magnetic compass. Then the following questions are posed:

> What do you think is happening to make that nail into a magnet?
>
> Does an electromagnet use electricity? What is electricity?
>
> Now put the compass near the wire, but away from the nail. What happens to the needle? Does it move? Why do you think it moves even when it is away from the nail?
>
> Do you think there is any way we could make a stronger electromagnet so we could pick up all this weight? How? Why would that work?
>
> If we used two batteries or a bigger battery, would that make a stronger electromagnet? Why?
>
> What do you think would happen it we wrapped the nail with twice as many wraps of wire? Do you think that would make the electromagnet weaker, stronger, or would it stay the same? Try it. What happened? Why?
>
> What makes an electromagnet work?

Equivocal answers to any question were followed up with specific probes.

In this way, the children were given maximum opportunity to explain the phenomena, and the testers were able to form a valid judgment of the child's theory underlying that explanation. Krupa, Selman, & Jaquette were able to classify the explanations into four levels of sophistication:

(1) *Perceptual Linkage*: overt properties, e.g., "The wires make it work." Ages 3 to 7.

(2) *Simplified Unseen Force*: covert, inanimate properties, e.g., "The power from the battery makes the compass needle move." Ages 5 to 9.

(3) *Coordinated Unseen Forces*: covert, multiple inanimate properties, e.g., "The electricity from the battery supplies power to make the magnet do things." Ages

7 to 15.

(4) *Balanced Systems*: transformations performed on the covert, coordinated forces, e.g., "The wire wrapped around the nail converts the electric energy into magnetic energy, and that is what causes the effect on the compass." Ages 12 to 18.

The same set of levels could be used to classify the children's explanations of the gravity demonstration. Students construct theories, theories that are more incomplete than wrong, theories that change in a systematic sequence as the child develops.

Although children must be tested individually and their responses classified by trained investigators, these results do indicate that it is possible to standardize the interview, and that a student's level of scientific understanding can be evaluated by asking for explanations.

Social Role of Explanations. Attribution theory is a central strand of social psychology and far too large a topic to be reviewed here. Suffice it to say that people are believed to be constantly accounting for what others do by attributing causal properties to them or to the environment. Causal attributions are arrived at by noticing covariation of events, and by determining whether the covariation is over people, time, or entities.

Abelson (1981) has pointed out that an account in terms of reasons, rather than causes, can be constructed from Schank & Abelson's (1977) theory of scripts. In terms of the psychology of explanation, the important thing about scripts is their stereotypy: once a script is invoked, the standard details can be filled in by default. On this view, scripts are a central means of organizing one's understanding of social interaction—they represent consensual knowledge of the behaviors appropriate in familiar situations. Seeking an explanation for scripted behavior, therefore, is unusual. Explanations are generally needed when events depart from the expected pattern (Lalljee & Abelson, 1983).

Lalljee, Watson, & White (1982) had college students judge situations as familiar or unfamiliar (e.g, supermarket or prison), and activities as expected or unexpected (e.g., talking or cheering). From these judgments they constructed such sentences as:

Gerry is talking in the supermarket.

John is talking in the prison.

Henry is cheering in the supermarket.

Walter is cheering in the prison.

Each sentence was printed on a sheet where space had been left for subjects to write a short essay explaining the event in their own words, and they were asked to rate their confidence in their explanation on a 7-point rating scale. Finally, subjects were asked whether the event had been caused more by the person or by the situation.

Unexpected behavior was was found to be attributed more to the person and expected behavior more to the situation, a result that they had expected. However, the explanations offered for unexpected behavior were more complex, so the comparison is not a simple one. Of interest in the present context, however, is the fact that subjects had no difficulty producing explanations, even for unexpected actions in unfamiliar situations. For most people, answering why-questions poses no problem.

Psychometric Studies of Explanation

J. P. Guilford's (1967) Structure-of-Intellect theory of intelligence proposes that individual differences in productive thinking can be analyzed into differences in convergent productions and in divergent productions. A test of convergent productions provides the information required for a subject to convege logically on unique answers; a test of divergent productions poses questions that may have many different answers. Clearly, the ability to construct explanations would be an instance of divergent productions, since many alternative explanations may be consistent with the available information.

Guilford discusses three tests of divergent productions:

Consequences: Asks for the consequences of some given event, e.g., people stop needing sleep.

Plot titles: Summarizes the plot of a story and asks for original, relevant titles for the story.

Uses test: Asks for all possible uses for some common object, e.g., a brick, a pencil, or a wire coathanger.

Guilford distinguishes two factors on the basis of responses to these tests. Divergent Semantic Units (DSU), generally known as the flexibility factor, is indicated by the number of obvious consequences, the number of ordinary plot titles, and the number of relevant uses. Divergent Semantic Transformations (DST), generally known as the originality factor, is indicated by remote consequences, clever plot titles, and unusual uses.

Tests of divergent productions have been used by some psychologists to study creativity. For example, Stratton & Brown (1972) used the plot-title problem to evaluate the effects of coaching on creative thinking. They found that if they coached subjects to increase their productivity, it worked, but it lowered the quality of their productions. Conversely, coaching on quality was also effective, but lowered productivity. Fortunately, coaching for both productivity and quality increased both.

It would not be difficult to develop an "Explanations Test" of divergent productions, similar to the three above, and to score it for flexibility and originality. The nearest thing to such a test seems to have been developed by Norman Frederiksen.

In 1959 Frederiksen created a performance test that simulated a problem commonly faced by scientists. He called it Formulating Hypotheses, on the assumption that searching for ideas, hypotheses, or possible solutions is a first step in solving any scientific problem. Each item consists of a graph or table showing findings from an actual research study. The task is to propose hypotheses to explain the findings.

For example, the sample item at the beginning of the test has a graph showing yearly rates of death from infectious diseases and from diseases of old age. The finding, printed below the graph, says: "Rate of death from infectious diseases has decreased markedly since 1900, while rate of death from diseases of old age has increased." The subject is provided with an answer sheet and asked to write "short statements of hypotheses (possible explanations) which you think might account for, or help to account for, the finding." A variety of different explanations can be proposed, of course.

The scoring procedure was developed inductively. First, a categorization was made of the ideas produced for each problem by a group of subjects. Then a definition of each category was written and a panel of judges evaluated the categories in the light of the information that was available to the subjects. Finally, a quality value was assigned to each category based on those evaluations. The scorer's task was merely to match each of the subject's responses to one of the response categories. From that information it was then possible to calculate the number of hypotheses generated, their average quality, and the number of hypotheses that were of high quality.

Formulating Hypotheses was used in a number of studies. Klein, Frederiksen, & Evans (1969) used it to test the the effect of giving feedback when the subject produced good explanations. Training increased the number of hypotheses, but not their quality. They also gave their subjects the Consequences test, but observed no transfer of training from Formulating Hypotheses to Consequences. Frederiksen & Evans (1974) repeated the study, but instead of coaching they allowed the experimental subjects to see some examples where many hypotheses were given, or where only very good hypotheses were given. In this case, the subjects' responses did move in the expected directions.

In 1978 Frederiksen & Ward developed four new tests that used a similar format. One them was again called Formulating Hypotheses: a graph is presented and hypotheses to explain it are requested. Another was Evaluating Proposals: the subject is asked to write critical comments about the methodology and design of several research proposals. In Solving Methodological Problems, each item is a brief statement of a problem encountered in planning a research study, and the task is to write suggested solutions. The fourth test was called Measuring Constructs, in which the subject is asked to suggest methods to elicit behavior relevant to psychological constructs like *leadership* or *bigotry* so that they can be measured objectively. All four tests pose problems of a kind that are likely to be encountered by research psychologists.

The scoring system was again developed by classifying responses, defining categories, and assigning quality values to the categories on the basis of expert

judgment. The scores that could be generated included the number of ideas, the number of unusual ideas, and the number of ideas that were both unusual and of high quality.

Ward, Frederiksen, and Carlson (1980) compared scores on these four tests with scores on the Graduate Record Examination (GRE). They found that quality scores were moderately related to GRE scores, but scores based on the number and unusualness of responses showed little relation to GRE scores. A follow-up study showed that the GRE tests were the best predictors of first-year grades in graduate school, but scores based on number and unusualness of the ideas were better predictors of such accomplishments as doing independent research, being author or co-author on a paper, and designing laboratory equipment. That is to say, tests of scientific thinking predicted real-life behavioral criteria better than did conventional aptitude tests.

Discussion

Frederiksen (1986) has described a procedure for developing and validating performance tests of the sort just considered, using as an illustration a study concerned with the selection of medical school students, where the criterion was a measure of clinical problem-solving ability. "The first step in the suggested procedure," he writes, "is to develop a theory about the criterion performance. A useful theory might best take the form of a model of the cognitive processes involved in the criterion behavior, based on some sort of task analysis" (p. 4).

The need for a good starting point in some theory of cognitive processes has been noted repeatedly by psychometricians and cognitive psychologists alike, yet anything more than a superficial matching of theoretical processes to measures of individual differences has proved very difficult. It is necessary to note, therefore, that a considerable gap still exists between the theoretical accounts of explanation and the practical testing of hypothesis formation that have been discussed in this working paper. To say, for example, that a physician needs to be able to formulate hypotheses that can explain a patient's symptoms is probably true, but it is hardly a theoretical insight deriving from current theories of cognitive processes. Moreover, to conclude from that

task analysis that a test battery for selecting medical students should include a test weighted on the flexibility factor is to move even further from current theories of cognitive processes.

The reconciliation of theories about processes with measures of individual differences is certainly not impossible in principle. One line worth exploring might begin with the theoretically motivated distinction between solving problems and constructing explanations (Wilensky, 1983, quoted above).

The present review of empirical studies of explantory skills does not include any studies using the "think-aloud" technique to investigate mathematical problem solving. Many such studies have been reported, and since thinking aloud requires a thinker to explain to the experimenter what he or she is doing, it might seem an ideal source of information about the processes of explanation. In fact, it is not. Someone trying to solve a mathematical problem is not trying to find explanations. Given a possible solution, the thinker may then try to verify its correctness, in which case explanatory skills may become involved. But searching a solution space and searching an explanation space seem to be as different as running forward and running backward are.

How might this difference be reflected in measures of individual differences? Guilford's distinction between convergent and divergent productions comes to mind: problem solving converges on a unique answer, whereas explanatory search produces divergent hypotheses. If the psychometric distinction betwen convergent and divergent productions is valid, perhaps it could be rationalized in terms of the theoretical distinction between forward chaining (in problem solving) and backward chaining (in explanation).

Whatever the merits of this suggestion, it is important to remember that not all thought processes are explanatory in purpose. An assessment of educational progress that relied exclusively on tests of explanatory skills would be as one-sided as an assessment that relied exclusively on tests of recognition memory. What is being recommended here is a variety of measuring instruments, where the evaluation of explanatory skills would be merely one component.

Educated people have complicated minds that require comparably complicated techniques of assessment.

Conclusion

The development and use of tests that sample students' ability to offer scientific explanations could add an important dimension to NAEP, one of theoretical and practical importance. The groundwork for developing such tests has already been laid; the results should provide data valuable for futher theoretical progress. Not the least, such tests could inspire teachers to devote more class time to the kind of cognitive processes involved in creating, evaluating, and generalizing scientific explanations.

References

Abelson, R. P. (1981). Psychological status of the script concept. *American Psychologist*, 36, 715-729.

Bromberger, S. (1966). "Why-questions," in Colodny, R. G. (Ed.), *Mind and Cosmos*. Pittsburgh: University of Pittsburgh Press.

Carey, S. (1985). *Conceptual Change in Childhood*. Cambridge, MA: MIT Press.

Chipman, S. F., Segal, J. W., & Glaser, R. (1985). *Thinking and Learning Skills, Volume 2: Research and Open Questions*. Hillsdale, NJ: Lawrence Erlbaum Associates.

DeJong, G. F. (1981). Generalizations based on explanations. *Proceedings of the 7th International Joint Conference on Artificial Intelligence*, Vancouver, B.C. Los Altos, CA: Morgan Kaufmann. Pp. 67-69.

DeJong, G. F., & Mooney, R. J. (1986). Explanation-based learning: An alternative view. *Machine Learning*, 1, (in press).

Frederiksen, N. (1959). *Development of the Test "Formulating Hypotheses": A Progress Report*. Princeton, NJ: Educational Testing Service.

Frederiksen, N. (1984). The real test bias: Influences of testing on teaching and learning. *American Psychologist*, 39, 193-202.

Frederiksen, N. (1986). Construct validity and construct similarity: Methods for use in test development and test validation. *Multivariate Behavioral Research*, 21, 3-28.

Frederiksen, N., & Evans, F. R. (1974). Effects of models of creative performance on ability to formulate hypotheses. *Journal of Educational Psychology*, 66, 67-82.

Frederiksen, N., & Ward, W. C. (1978). Measures for the study of creativity in scientific problem solving. *Applied Psychological Measurement*, 2, 1-24.

Glaser, R. (1985). Preface. In Segal, J. W., Chipman, S. F., and Glaser, R. (Eds.), *Thinking and Learning Skills, Volume 1: Relating Instruction to Research*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Goodman, N. (1955). *Fact, Fiction, and Forecast*. Cambridge, MA: Harvard University Press.

Green, B. F. A primer of testing. *American Psychologist*, 36, 1001-1011.

Guilford, J. P. (1967). *The Nature of Human Intelligence*. New York: McGraw-Hill.

Hammond, K. J. (1986). Learning to anticipate and avoid planning problems through the explanation of failures. *Proceedings of the 5th National Conference on Artificial Intelligence*, Philadelphia, PA. Los Altos: Morgan Kaufmann. Vol. 1, pp. 556-560.

Harman, G. (1986). *Change in View: Principles of Reasoning*. Cambridge, Mass.: MIT Press.

Harman, G., Cullingford, R., Bienkowski, M., Salem, K., & Pratt, I. (1986). Default defeaters in explanation-based reasoning. Technical Report No. 3, Cognitive Science Laboratory, Princeton University.

Hoffmann, B. (1962). *The Tyranny of Testing*. New York: Crowell-Collier.

Kass, A. (1986). Modifying explanations to understand stories. *Eighth Annual Conference of the Cognitive Science Society*, Amherst, MA. Hillsdale, NJ: Lawrence Erlbaum Associates. Pp. 691-696.

Kedar-Cabelli, S. T. (1985). Purpose-directed analogy. *Proceedings of the Cognitive Science Society Conference*. Irvine, CA: Morgan Kaufmann.

Klein, S. P., Frederiksen, N., & Evans, F. R. (1969). Anxiety and learning to formulate hypotheses. *Journal of Educational Psychology*, 60, 465-475.

Krupa, M. P., Selman, R. L., & Jaquette, D. S. (1985). The development of science explanations in children and adolescents: A structural approach. In Chipman, S. F., Segal, J. W., & Glaser, R. (Eds.), *Thinking and Learning Skills, Volume 2: Research and Open Questions*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Lalljee, M., & Abelson, R. P. (1983). The organization of explanations. In Hewstone, M. (Ed.), *Attribution Theory: Social and Functional Extensions*. Oxford: Basil Blackwell.

Lalljee, M., Watson, M., & White, P. (1982). Explanations, attributions and the social context of unexpected behavior. *European Journal of Social Psychology*, 12, 17-29.

Leake, D. B., & Owens, C. C. (1986). Organizing memory for explanation. *Eighth Annual Conference of the Cognitive Science Society*, Amherst, MA. Hillsdale, NJ: Lawrence Erlbaum Associates. Pp. 710-715.

Lewis, C. H. (1986). Understanding what's happening in system interactions. In Norman, D. A., & Draper, S. W. (Eds.), *User Centered System Design: New Perspectives on Human-Computer Interaction*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Lewis, C. H., & Mack, R. L. (1982). Learning to use a text processing system: Evidence from "thinking aloud" protocols. *Proceedings of the Conference on Human Factors in Computer Systems*. New York: ACM, 387-392.

Mack, R. L., Lewis, C. H., & Carroll, J. M. (1983). Learning to use word processors: Problems and prospects. *ACM Transactions on Office Information Systems*, 1, 254-271.

Miller, G. A. (1979). Images and models, similes and metaphors. In Ortony, A. (Ed.), *Metaphor and Thought*. New York: Cambridge University Press.

Mitchell, T. M., Keller, R. M., & Kedar-Cabelli, S. T. (1986). Explanation-based generalization: A unifying view. *Machine Learning*, 1, 47-80.

Mooney, R. J., & Bennett, S. W. (1986). A domain independent explanation-based generalizer. *Proceedings of the 5th National Conference on Artificial Intelligence*, Philadelphia, PA. Los Altos: Morgan Kaufmann. Pp. 551-555.

Mooney, R. J., & DeJong, G. F. (1985). Learning schemata for natural language

processing. *Proceedings of the 9th International Joint Conference on Artificial Intelligence*, Los Angeles, CA. Los Altos, CA: Morgan Kaufmann. Pp. 681-687.

Nagel, E. (1961). *The Structure of Science: Problems in the Logic of Scientific Explanation*. New York: Harcourt, Brace & World.

Newell, A., & Simon, H. A. (1972). *Human Problem Solving*. Englewood Cliffs, NJ: Prentice-Hall.

Nickerson, R. S., Perkinds, D. N., & Smith, E. E. (1985). *The Teaching of Thinking*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Piaget, J. (1929). *The Child's Conception of the World*. London: Routledge and Kegan Paul.

Piaget, J. (1960). *The Child's Conception of Physical Causality*. Totowa, NJ: Littlefield, Adams.

Polya, G. (1957). *How to Solve It*. (2nd edition). Princeton: Princeton University Press; New York: Doubleday.

Resnick, L. B. (1986). *Education and Learning to Think*. A special report prepared for the Commission on Behavioral and Social Sciences and Education. Washington, DC: National Research Council.

Schank, R. C. (1981). Language and memory. In Norman, D. A. (Ed.), *Perspectives on Cognitive Science*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Schank, R. C. (1982). *Dynamic Memory: A Theory of Learning in Computers and People*. New York: Cambridge University Press.

Schank, R. C. (1986). Explanation: A first pass. In Kolodner, J. L., & Riesbeck, C. K. (Eds.), *Experience, Memory, and Reasoning*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Schank, R. C. (In press). *Explanation Patterns: Understanding Mechanically and Creatively*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Schank, R. C., & Abelson, R. (1977). *Scripts, Plans, Goals, and Understanding*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Schoenfeld, A. H. (1985). *Mathematical Problem Solving*. New York: Academic Press.

Segal, J. W., Chipman, S. F., & Glaser, R. (1985). *Thinking and Learning Skills,*

Volume 1: Relating Instruction to Research. Hillsdale, NJ: Lawrence Erlbaum Associates.

Simon, H. A. (1980). Problem solving and education. In Tuma, D. T., & Reif, F. (Eds.), *Problem Solving and Education: Issues in Teaching and Research.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Stratton, R. P., & Brown, R. (1972). Improving creative thinking by training in the production and/or judgment of solutions. *Journal of Educational Psychology*, 63, 390-397.

Ward, W. C., Frederiksen, N., & Carlson, S. B. (1980). Construct validity of free-response and machine scorable tests. *Journal of Educational Measurement*, 17, 11-29.

Wigdor, A. K., & Green, B. F., Jr. (1986). *Assessing the Performance of Enlisted Personnel: Evaluation of a Joint-Service Research Project.* National Research Council. Washington, DC: National Academy Press.

Wilensky, R. (1983). *Planning and Understanding: A Computational Approach to Human Reasoning.* Reading, MA: Addison-Wesley.